SCHRAUSSER

Thesis chapter 1: Introduction (part II, permutation tests for repeated measurement designs)

Christian Scambor & Dietmar G. Schrausser

Institute of Psychology, Karl Franzens University, Universitätsplatz 2, 8010 Graz, Austria

Thesis[1] April 1997; Slightly revised english translation January 2023

Abstract

[…][2] The advantages of permutation methods over approximate methods (*AP*, methods based on distribution assumptions) are explained in Chapter 1.2 and worked out in Chapter 1.3. It is shown how *confounding variables* are controlled (Chap. 1.4). In Chapter 1.5 experiments with *repeated measurement* designs are dealt with.

1.1.3 Effect sizes ε

As with *AP* methods, effect sizes ε can also be taken into account in permutation tests. If one assumes that the difference between two treatments *A* and *B* would be reflected, for example, in a mean difference of magnitude *d*, where $A > B$, then an amount *d* is added to each measurement value of group *B* before the permutation test is carried out. The null hypothesis H_0 states that there is no difference between the two experimental groups from which the measurements of one group were transformed.

1.1.4 Number of permutations *M* and exact *p*-value

The number of permutations *M* for an *unpaired* design with *N* subjects and *k* experimental conditions, each assigned n_1, n_2, …, n_k subjects, is calculated[3] according to

(1.2)
$$M = \frac{N!}{n_1!\, n_2! \ldots n_k!}$$

[1] Scambor, C. (1997). Permutationsverfahren in Einzelfallstudien: Grundlagen, Anwendungen und Teststärke (Permutation methods in single case studies: Foundations, applications and power). *Thesis*. Institute of Psychology, Karl Franzens University, Austria.

[2] The introductory presentation of the permutation test in the *unpaired* 2-sample problem is omitted here. The reader will find in-depth elaborations on this, e.g. in Schrausser & Scambor (1995, 2022) or Schrausser (1996, 1998 and 2022a).

[3] Intra-condition permutations (permutations in which the order of the values *within* a treatment is changed) are not taken into account, since they leave every meaningful test-value θ uneffected.

It is important to consider the number of permutations when designing the experiment, as this number sets the smallest achievable *p*-value. By definition, this exact *p*-value is calculated using the following formula:

(1.2.1)
$$p = M^{-1} \sum_{i=1}^{M} \theta_i \geq \Theta$$

with
M: Number of possible permutations for a given design;
θ_i : Characteristic- or test-value *theta* of permuted design arrangement *i*;
Θ: Characteristic- or test-value *Theta* from the initial, experimental conditon.

Furthermore, the test power *1-β* (cf. Chapter 3.1.3) of a procedure depends on the *number of permutations*. In general, the following applies: the more measuring points, the more permutations are possible, the higher the power of this procedure (restrictions on this *rule of thumb*: cf. Chap. 7.6.3).

As *N* increases, the number of permutations *M* increases rapidly, so that it can be impractical to calculate the combinatorially complete number of permutations, even when using fast computers. If only a random sample of permutations is used to calculate the *p*-value, the term *randomized permutation test* (also: *Monte Carlo version*) is used, if the complete *M* of permutations is used, it is called the *systematic or exact permutation test* (cf. Schrausser, 1996).

1.2 Advantages of permutation tests

Edgington (1995) emphasizes the following advantages of permutation tests:

1.2.1 Intelligibility

Permutation tests are generally easy to understand. Inferential statistical considerations (cf. Chapter 3.1) are dispensed with and only the available data are considered. Values Θ that result from the observed raw data are extreme or are not, relative to the other possible values θ. This forms the basis for the decision about the H_0.

1.2.2 Flexibility

At the level of the parameters θ, permutation tests are flexible because the examiner can choose a parameter adapted to the experiment and determine it himself. This means that it is not necessary to fall back on conventional test-values such as *t* or *F*, but one can use e.g. simple sumscores Σ instead (to compare two groups). This makes statistical calculations easier to understand.

At the level of application, permutation tests are flexible because they open up areas that are difficult or impossible to access using conventional statistical methods (e.g. single-case experiments).

1.2.3 Few prerequisites

Assumptions of permutation tests are limited to (1) experimental independence and (2) the presence of a random element.

ad (1) experimental independence is defined as follows:

> Two subjects are experimentally independent if one subject does not influence the measurement of the other subject. (Edgington, 1995, S.14)

This requirement applies to any experiment, regardless of how the results are statistically evaluated. Experimental independence is violated, for example, if the behavior (speaking, expressions of dissatisfaction, etc.) of individual subjects influences the behavior of other subjects in a group experiment in which several subjects are examined at the same time. In the group experiment, certain measures, such as isolating people from each other, can be used to ensure experimental independence.

Ad (2) Permutation tests can be used if the random element is realized either by (a) *drawing* random samples or by (b) randomly *allocating* subjects to treatments. The second option makes it possible to circumvent unrealistic requirements of *AP* methods (cf. Chapter 1.3) when carrying out an experiment. Only the available data are used to determine the significance level *p*, estimates of population parameters are omitted.

1.2.4 Validity

Permutation tests are *exact* tests whose results are valid for the examined sample. In contrast to this, the use of *AP* methods in psychology often leads to violations of assumptions, especially the assumption of drawing *random* samples from populations (cf. Chapter 1.3). The *validity of the statistical conclusion* (cf. Bredenkamp, 1980, p. 14; Hager, 1987, p. 113 ff.) deals with the question of whether an appropriate statistical procedure was chosen. *AP* methods do not deliver valid results if the random sampling requirement is violated, or are only valid to the extent that they do not deviate too much from the results of the permutation tests. In this context, permutation tests can be considered as a *validity criterion* for *AP* procedures (cf. Bredenkamp, 1980, p. 21; Edgington, 1995, p.10 ff.).

1.3 The random sampling problem

AP methods are inference-oriented (see Chapter 3), i.e. they are designed to make conclusions about populations from which random samples are drawn. This approach has often been criticized (e.g. Edgington, 1995, p. 6 ff.; Bredenkamp, 1980, p. 18 f.). The main point of criticism is the practical impossibility in psychology of drawing random samples. In Chapter 1.3.1 the random sample is defined. In Chapter 1.3.2 it is shown that the assumption of random sampling in psychology is unrealistic and what kind of violations to this assumption occur. In Chapter 1.3.3 the argument that representative samples could be used instead of random samples is rejected. Chapter 1.3.4 deals with the role of robustness studies in context with the violation of the random sampling assumption. An alternative approach is offered in Chapter 1.3.5. The problem of the generalizability of sample results is dealt with in Chapter 1.3.6.

1.3.1 Definition

Random samples are defined as follows:

> Eine zufällige Auswahl von Untersuchungseinheiten aus einer Grundgesamtheit bezeichnen wir als eine Zufallsstichprobe. Sie ist dadurch gekennzeichnet, daß jedes Element der Grundgesamtheit, unabhängig davon, welche weiteren Elemente schon zur Stichprobe gehören, mit gleicher Wahrscheinlichkeit ausgewählt werden kann. (Bortz, 1989, S.113)

> We refer to a random selection of examination units from a population as a random sample. It is characterized in that each element of the population, regardless of which other elements already belong to the sample, can be selected with the same probability. (Bortz, 1989, p .113)

Schrausser (1996, p. 2-1 ff.; 2022b) provides a heuristic example that clearly illustrates the mechanism of random selection.

1.3.2 The random sample in psychological research; violations

It is obvious that such a procedure is practically impossible in psychology. Psychological hypotheses often refer to an overall population of a totality of people, i.e. to an open population (cf. Edgington, 1995, p.6 ff.; Bredenkamp, 1980, p.19). Populations may be definable in closed form in survey research; in this case, random samples could then be drawn and *AP* methods could justifiably be used. In most psychological studies, however, this is not the case; instead, the following violations of the random sample requirement occur (cf. Edgington, 1995, p. 6 ff.): (1) an *accidental*[4] sample (i.e. every *available* subject) is examined; (2) there is a *systematic* selection of subjects (e.g. voluntary); (3) populations are *inferred* from subpopulations; even if random samples were drawn from the

[4] Non probability sample *not* to be confused with the random sample.

subpopulation, such a conclusion is invalid. The procedure is then usually as follows: The sample examined (usually an accidental sample) is *treated* as a random sample. However, a warning is given against this procedure:

> Man begnügt sich [...] gelegentlich mit sog. *anfallenden* oder *ad hoc* Stichproben [...] in der Hoffnung, auch so zu aussagefähigen Resultaten zu gelangen. Vor dieser Vorgehensweise sei nachdrücklich gewarnt. Zwar ist die Verwendung inferenzstatistischer Verfahren nicht daran gebunden, daß eine Stichprobe zufällig aus einer wirklich existierenden Population gezogen wird; letztlich läßt sich für jede *Stichprobe* eine fiktive Population konstruieren, für die diese *Stichprobe* zufällig erscheinen mag. Die Schlüsse, die aus derartigen Untersuchungen gezogen werden, beziehen sich jedoch auf nicht real existierende Populationen und können daher wertlos sein. (Bortz, 1993, S. 85)

> One is satisfied [...] occasionally with so-called *accidental* or *ad hoc* samples [...] in the hope of obtaining meaningful results in this way. We expressly warn against this approach. It is true that the use of inferential statistical methods is not involved bound to a sample being randomly drawn from an actual population; ultimately, for any *sample* a fictitious population can be constructed for which that *sample* may appear random, but the conclusions drawn from such studies are related to populations that do not actually exist and can therefore be worthless. (Bortz, 1993, p. 85)

1.3.3 Representativeness

The use of *AP* methods is often justified by referring to the representativeness of the samples (cf. Selg & Bauer, 1976, p.87; Bortz, 1989, p.113). A sample is considered representative for a population if the characteristics considered relevant for the study have similar distributions in the population *and* in the sample. Edgington (1995, p. 6) does not accept this argument: *AP* methods *require* random sampling; this requirement is also violated if a sample can be considered representative but was *not randomly* selected.

1.3.4 Robustness studies

Robustness studies examine the behavior of statistical methods when various assumptions are violated. Questions that robustness studies try to answer can be formulated e.g. as follows:

> How much would the use of our standard significance tables mislead us if the populations that were randomly sampled had quite different variances instead of the identical variances that are assumed by the tables? (Edgington, 1995, S. 13)

If one considers the use of *AP* methods in question, such investigations are of no help, since they deal with population characteristics instead of sample characteristics (cf. Edgington, 1995, p. 12 f.).

1.3.5 Permutation tests for random sampling and random allocation; Classification scheme for permutation tests

Permutation tests can be used when a random sample is actually available. The abbreviation *P Tests* stands for these methods (cf. Schrausser, 1996). The great advantage of these methods, however, is that they deliver valid results for any sample, regardless of how this sample was compiled. The only condition is that subjects be at least *randomly assigned* to treatments. In this case we speak of *permutation tests in the randomization model* or *randomization tests* (*PR tests*).

> Determining significance on the basis of a distribution of test statistics generated by permuting the data is characteristic of all permutation tests; it is when the basis for permuting the data is random assignment that a permutation test is called a randomization test. (Edgington, 1995, S. 1 f.)

With respect to the explanations in Chapter 1.1.4, permutation tests can therefore be classified (1) with regard to the random element (random sample or random allocation) and (2) with regard to the inclusion of either (a) every possible or (b) only a part of possible permutations in the calculation of the exact *p*-value. For the classification scheme of permutation (exact) tests s. Schrausser (1996, p. 1-14).

Since the present work also discusses cases in which even the random allocation is replaced by assumptions (cf. Chapter 4), the more general term *permutation tests* is used. Permutation tests in the presence of a *random* sample are not discussed in the further explanations.

1.3.6 Generalizability

The great advantage of *AP* methods is that conclusions can be drawn about population ratios from sample results. However, as has been shown, this approach is rarely justified. Therefore, permutation tests represent a good alternative to *AP* methods. However, the results from permutation tests are only valid for the examined sample and do not say anything about population ratios. Regarding the problem of the generalizability of such results, Edgington says:

> Inferences about treatment effects for other subjects must be nonstatistical inferences - inferences without a basis in probability. Nonstatistical generalization is a standard scientific procedure. We generalize from our experimental subjects to individuals who are quite similar in those characteristics that we consider relevant. (Edgington, 1995, S. 8)

A different approach (*deductivistic theory of experiment*, cf. Bredenkamp, 1980; 1996; Erdfelder & Bredenkamp, 1994) is dealt with in Chapters 2 and 3.

1.4 Generalization of the principle and control of confounding variables

First, a distinction is made between person-related and operational confounding variables (Chap. 1.4.1). The principle of permutation tests is generalized in Chapter 1.4.2; general terms are introduced and it is explained how to control for person-related confounding variables. Chapter 1.4.3 deals with complex experimental units that are used to control operational confounding variables.

1.4.1 Operational and person-related confounding variables

With an experiment, a distinction is made between (a) operational and (b) person-related confounding variables. operational confounding variables are countered by control measures (switching off, keeping constant, registering, etc.) (cf. Bortz, 1984, p.400 ff.).

The control of person-related confounding variables is achieved by randomly assigning subjects to treatments (*randomization*, cf. e.g. Bredenkamp, 1980, p.1). This makes the samples comparable; the likelihood that the groups differ on the dependent variable (or on other variables that may affect the dependent variable) decreases. This control measure also serves this purpose when experimental results are evaluated with permutation tests:

> The random assignment of subjects to treatments is to control for variability among subjects with regard to subject variables, i.e., individual differences. (Edgington, 1995, S.26)

1.4.2 Generalization; Controlling for person-related confounding variables

The principle of randomization tests is that subjects are randomly assigned to treatments e.g. *A* and *B*. The random assignment ensures that the subjects in the groups do not differ: the groups are made *comparable*. Random allocation is a control measure for person-related confounding variables. The general formulation *experimental unit* refers to units that are assigned to treatments. In the simplest case, these are people or points in time (as in the single case experiment, cf. Chapter 7). If several treatments are applied to one person, several times are set for this. Then these points in time are *randomly* assigned to the treatments. This measure controls the intra-individual variability of a person's behavior over time (exercise effects, fatigue effects, etc.).

1.4.3 Complex experimental units; Control of operational confounding variables

Complex experimental units represent a special case. Two types of complex experimental units have been defined (Edgington, 1995, p. 27 f.):

The experimental unit does not consist of a single person but of a (1) *group of subjects* (e.g. school classes as experimental units); introduced.

Items (2) *other than* subjects or time points are included in the experimental unit. This is for the purpose of controlling for operational confounding variables. For example, in an experiment conducted by two investigators, the influence of the investigator on the respondent's response could be controlled

as follows: Half of the subjects are assigned to investigator 1 and the other half to investigator 2. An *experimental unit thus consists of* the *test person and an investigator*. The complex experimental units formed in this way are randomly assigned to treatments *A* and *B*. Any number of confounding variables can be controlled using this technique.

The same technique is used to control for confounding variables in *single case* experiments. It should only be noted that initially a *point in time* represents the experimental unit. Each of the time points is assigned to one of the two investigators, randomly or arbitrarily. Each complex experimental unit (consisting of *time point and investigator*) is then randomly assigned to a treatment.

1.5 Repeated measurement

The permutation test procedure for experiments with repeated measurements is to be explained using an example where every subject *i* receives both treatments *A* and *B*. The inter-subject variability is not a problem in this case. However, it is necessary to control for the intra-subject variability. The procedure is as follows: Each of the *n* persons receives *k* treatments. For each person there are therefore *k* points in time at which a treatment is applied. An experimental unit is defined as person × time point. Overall there are $N = n \times k$ experimental units. The experimental units for one subject are randomly assigned to the *k* treatments. The allocation for each individual is *independent* of the allocation for the other individuals. There are *M* permutations *P* possible, where

(1.3) $$M = (k!)^n$$

with
k: Number experimental units, treatments;
n: Number of subjects.

In Fig. 1.1 a general data scheme (*experiment plan*, cf. Hager, 1987, p. 56 ff.) and characteristic values θ that are used for permutation tests in experiments with *repeated measurements* are shown.

In following Example 1.2 the procedure for experiments with repeated measurements is explained using simulated raw values *x* in accordance with the representations in (Fig. 1.1).

Figure 1.1

		Treatment					
		1	2	···	*j*	···	*k*
Subject	1	x_{11}	···	···	x_{1j}	···	x_{1k}
	2	:			:		:
	:	:			:		:
	i	x_{i1}	···	···	x_{ij}	···	x_{ik}
	:	:			:		:
	n	x_{n1}	···	···	x_{nj}	···	x_{nk}
		T_1	···	···	T_j	···	T_k

with
x_{ij}: Raw value *x* of subject *i* under treatment *j*;
T_j: θ, Sum of raw values *x* of a treatment *j*.

Example 1.2

Assume $n = 3$ subjects ($i = 1, 2, 3$) receive $k = 2$ treatments (*A*, *B*; $k = 1, 2$) each. Two points in time are determined for each person, which together with the person form the *experimental units* that are randomly assigned to the treatments. Some of the test persons therefore receive the treatments in the order *A* - *B*, some in the order *B* - *A*. A *t*-value is conceivable as a parameter θ - analogous to the *AP*

method of the *t* test for dependent random samples. Here, following *equivalent* characteristic values θ are easier to calculate:

For the *two-tailed* test, the *sum of the squared sums* of raw values *x* from both treatments (*SQMS*) is used:

(1.4) $$\sum_{j=1}^{k} T_j^2 = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} x_{ij} \right)^2 \rightarrow \sum_{j=1}^{2} T_j^2 = \left(\sum_{i=1}^{3} x_{i1} \right)^2 + \left(\sum_{i=1}^{3} x_{i2} \right)^2$$

with sum of raw values (*MS*) of a treatment:

(1.5) $$T_j = \sum_{i=1}^{n} x_{ij}$$

with x_{ij} raw value of person *i* under treatment *j*.

For the *one-tailed* test, the *sum* of the raw values *x* of that treatment for which *higher* values are predicted is used (e.g. treatment *B*, $j = 2$, T_2):

(1.6) $$T_2 = \sum_{i=1}^{3} x_{i2}$$

The results were as follows:

Table 1.4 Raw-values for Example 1.2

	Treatment k	
Subject n	*A*	*B*
1	1	4
2	2	6
3	2	7

Every possible raw data permutation P_M is shown in Table 1.5. The first *permutation* P_0 corresponds to the observed, initial data arrangement.

For the two-sided test, only for the last permutation P_7 a test value $\theta = 314$ is obtained, that is greater or equal to the characteristic value $\Theta = 314$ of the observed, initial data arrangement P_0. According to eq. 1.2.1 the two-tailed *p*-value is therefore $p^2 = {}^2/_8 = 0.250$, for the one-sided test, the observed value $\theta = 17$ of P_0 alone is the largest, so $p^1 = {}^1/_8 = 0.125$.

Table 1.5 Permutations P_M of the raw values *x* from Tab. 1.4 with $M = (k!)^n = 8$ and intermediate steps for calculating parameters θ, *T* and T^2 for obtaining exact one- and two-tailed *p*-values, $p^1 = (T \geq T_0)/_M$, $p^2 = (T^2 \geq T^2_0)/_M$, respectively.

	P													
	0		*1*		*2*		*3*		*4*		*5*		*6*		*7*	
n	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*
1	1	**4**	1	**4**	1	**4**	**4**	1	1	**4**	**4**	1	**4**	1	**4**	1
2	2	**6**	2	**6**	**6**	2	2	**6**	**6**	2	**6**	2	2	**6**	**6**	2
3	2	**7**	**7**	2	2	**7**	2	**7**	**7**	2	2	**7**	**7**	2	**7**	2
T	5	17	10	12	9	13	8	14	14	8	12	10	13	9	17	5
T^2	314		244		250		260		260		244		250		314	

In the case of repeated measurements, it should be noted that the raw values *x* per line belong to a single person. So, given the validity of the null hypothesis H_0 (*treatments have no influence on* the *AV*), raw values of each individual case can be permuted over the treatments, but raw values of one case *must* not be assigned to another case. For this reason, *only* ***line-by-line*** *permutations* are permitted. The

problem of carry-over effects, which are relevant for experiments with repeated measurements, is dealt with in context with single case experiments (cf. Chapter 7.6.1).

1.6 Summary

The basic operation of permutation tests was displayed. The advantages of this method were worked out and compared to the problems that arise when using *AP* methods. Permutation tests *P* represent a *realistic* possibility of *validating* results from experiments (and other types of investigation, cf. Chapter 4) in a random-critical manner. However, there are problems of a scientific-philosophical nature related to the decision between H_0 and H_1. For this reason, it is necessary to consider *AP* procedures and permutation tests from a scientific-philosophical perspective (cf. Chapters 2 and 3) in order to work out the possibilities as well as limitations of *both* approaches *AP* and *P*.

References

Bortz, J. (1984). *Lehrbuch der empirischen Forschung*. Berlin: Springer.

Bortz, J. (1989). *Statistik für Sozialwissenschaftler* (3. Aufl.). Berlin: Springer.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente*. Darmstadt: Steinkopff.

Bredenkamp, J. (1996). Grundlagen experimenteller Methoden. In: Erdfelder, E., Mausfeld, R., Meiser, T., Rudinger, G.: *Handbuch Quantitative Methoden*. Weinheim: Psychologie Verlags Union, S. 37-46.

Edgington, E. S. (1995). *Randomization tests* (3rd ed). New York: Marcel Dekker.

Erdfelder, E., & Bredenkamp, J. (1994). Hypothesenprüfung. In: Herrmann, T., Tack, W. H. (Hrsg.): *Methodologische Grundlagen der Psychologie. Enzyklopädie der Psychologie, Themenbereich B, Serie I*. Göttingen: Hogrefe, S. 604-648.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In: Lüer, G. (Hrsg.): *Allgemeine experimentelle Psychologie: Eine Einführung in die methodischen Grundlagen mit praktischen Übungen für das experimentelle Praktikum*. Stuttgart: Fischer, S. 43-264.

Schrausser, D. G. (1996). Permutationstests: Theoretische und praktische Arbeitsweise von Permutationsverfahren beim unverbundenen 2 Stichprobenproblem. *Thesis*. Institute of Psychology, Karl Franzens University, Austria. DOI: 10.13140/RG.2.2.24500.32640/1

Schrausser, D. G. (1998). Exakte Verfahren oder Asymptotische Approximation. In J. Glück. M. Jirasco & B. Rollett (Hrsg.). *Perspektiven psychologischer Forschung in Österreich - Teil 2*. Wiener Universitätsverlag.

Schrausser, D. G. (2022a) Thesis chapter 1: Introduction. In: Schrausser, D. G. (1996) *Permutationstests: Theoretische und praktische Arbeitsweise von Permutationsverfahren beim unverbundenen 2 Stichprobenproblem*. Thesis. Karl Franzens University, Institute of Psychology.

Schrausser, D. G. (2022b) Thesis chapter 2: Random selection as correlative construct. In: Schrausser, D. G. (1996) *Permutationstests: Theoretische und praktische Arbeitsweise von Permutationsverfahren beim unverbundenen 2 Stichprobenproblem*. Thesis. Karl Franzens University, Institute of Psychology.

Schrausser, D. G., & Scambor, C. (1995). Randomisierungstests. Eine Einführung in die Methode. *Manuskript*. Institut für Psychologie, Karl Franzens Universität, Graz.

Schrausser, D. G., & Scambor, C. (2022). Randomizationtests. An introduction to the Method. *Manuscript* (English translation). Karl Franzens University, Institute of Psychology.

Selg, H., & Bauer, W. (1976). *Forschungsmethoden der Psychologie* (3. Aufl.). Stuttgart: Kohlhammer.